JADAVPUR UNIVERSITY

MASTER DEGREE THESIS

Correlation between Network Properties and Structure of Online Social Networks

---

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Technology in Distributed and Mobile Computing

in the

School of Mobile Computing and Communication

by
SUBHAYAN BHATTACHARYA
University Roll Number: 001730501001
Examination Roll Number: M4DMC19005
Registration Number: 141102 of 2017-2018

Under the Guidance of
Dr. SARBANI ROY
Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032

May 21, 2019

Declaration of Authorship

I, Subhayan Bhattacharya, declare that this thesis titled, "Correlation between Network Properties and Structure of Online Social Networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a masters degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

To whom it may concern

This is to certify that Subhayan Bhattacharya has satisfactorily completed the work in this thesis entitled "Correlation between Network Properties and Structure of Online Social Networks", University Roll Number: 001730501001, Examination Roll Number: M4DMC19005, Registration Number: 141102 of 2017-2018. It is a bonafide piece of work carried out under my supervision at Jadavpur University, Kolkata-700032, for partial fulfillment of the requirements for the degree of Master of Technology in Distributed and Mobile Computing from the School of Mobile Computing and Communication, Jadavpur University for the academic session 2017-2019.

Dr. Sarbani Roy
Associate Professor
Department of Computer Science & Engineering,
Jadavpur University
Kolkata-700032.

Director
School of Mobile Computing and Communication,
Jadavpur University
Kolkata-700032.

Prof. Pankaj Kumar Roy
Dean, Faculty of Interdisciplinary Studies, Law and Management
Jadavpur University
Kolkata-700032.

Certificate of Approval

(Only in case the thesis is approved)

This is to certify that the thesis entitled "Correlation between Network Properties and Structure of Online Social Networks" is a bona-fide record of work carried out by Subhayan Bhattacharya, University Roll Number: 001730501001, Examination Roll Number: M4DMC19005, Registration Number: 141102 of 2017-2018, in partial fulfilment of the requirements for the award of the degree of Master of Technology in Distributed and Mobile Computing from the School of Mobile Computing and Communication, Jadavpur University for the academic session 2017-2019. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

_____

(Signature of the Examiner)
Date:

_____

(Signature of the Examiner)
Date:

Jadavpur University

Abstract

Faculty of Interdisciplinary Studies, Law and Management, Jadavpur University

School of Mobile Computing and Communication

Master of Technology in Distributed and Mobile Computing

Correlation between Network Properties and Structure of Online Social Networks

by

SUBHAYAN BHATTACHARYA
University Roll Number: 001730501001
Examination Roll Number: M4DMC19005
Registration Number: 141102 of 2017-2018

Online Social Networks(OSNs) are abundant and widely used in current times. It is used as a platform for communication among digital citizen. The frequency, method, and purpose of communication are different for different OSNs. While some OSNs have their users hooked to the screen for the better part of the day, other OSNs have users visiting once a day. While some OSN has small textual content for communication, some OSNs are dependent on audio-visual contents, and other OSNs have a mixture of both. While some OSNs are used mainly to keep in touch with friends, colleagues and relatives, other OSNs are used to meet new people. However, this variety in OSNs is not captured in literature. As a part of the Masters Degree Thesis, a categorisation of Online Social Networks based on network properties such as Centrality Measures, Clustering Coefficient, Homophily, Assortativity and so on has been presented that provides an insight into the network structure and behaviour based on the categorisation.

Acknowledgements

On the submission of "Correlation between Network Properties and Structure of Online Social Networks", I wish to express gratitude to the School of Mobile Computing and Communication for sanctioning a thesis work under Jadavpur University under which this work has been completed.

I want to convey my sincere gratitude to **Dr Sarbani Roy**, Associate Professor, Department of Computer Science & Engineering, Jadavpur University for her valuable suggestions throughout the project duration. I am grateful to her for her constant support which helped me a lot to fully involve myself in this project and develop new approaches in the field of Social Network Analysis.

I would like to express my sincere, heartfelt gratitude to Mrs Sankhamita Sinha, Assistant Professor, Meghnad Saha Institute of Technology, Kolkata, for suggestions and guidance.

I would also wish to thank Director of the School of Mobile Computing and Communication, Jadavpur University and Prof. Pankaj Kumar Roy, Dean, Faculty of Interdisciplinary Studies, Law and Management, Jadavpur University for providing me all the facilities and for their support to the activities of this research.

Lastly, I would like to thank all my teachers, classmates, guardians and well-wishers for encouraging and co-operating me throughout the development of this project. I would like to especially thank my parents whose blessings helped me to carry out my project in a dedicated way.

Regards,
SUBHAYAN BHATTACHARYA
University Roll Number: 001730501001
Examination Roll Number: M4DMC19005
Registration Number: 141102 of 2017-2018
School of Mobile Computing and Communication
Jadavpur University

Signed:

Date:

Contents

List of Figures

List of Abbreviations

OSN	**O**nline **S**ocial **N**etwork
PDF	**P**robability **D**ensity **F**unction
CDF	**C**umulative **D**ensity **F**unction
PLE	**P**ower **L**aw **E**xponent
GCC	**G**lobal **C**lustering **C**oefficient
LCC	**L**ocal **C**lustering **C**oefficient
ALCC	**A**verage **L**ocal **C**lustering **C**oefficient

List of Symbols

$d(i)$	degree distribution
D_{net}	Network Density
$\rho(i,j)$	distance between two nodes i and j
l	Average Path Length
γ	Power Law Exponent
gcc	Global Clustering Coefficient
$lcc(i)$	Local Clustering Coefficient of node i
$alcc$	Average Local Clustering Coefficient
$c(i)$	Closeness Centrality
$e(i)$	Eigenvector Centrality
$b(i)$	Betweenness Centrality
q_k	normalised distribution of the excess degree
σ	standard deviation of q_k
p_k	probability of degree k
r	Assortatvity
h	Homophily

Chapter 1

Introduction

1.1 Overview

An Online Social Network(OSN) is an online platform that allows it is members to see, connect, and communicate with other members. The content on an OSN might be textual, images, audio, video or a mixture of any of these. There may or may not be a profile for the members of the network. The members of an OSN can choose to connect with other members of the OSN and gain rights to view and interact with the other members. The mode of communication can be direct(personal messages) or in-direct(posts, comments, tags, replies, retweets and so on). The connection and communication are established and maintained online. There exist several definitions for OSNs, including the one provided by Boyd and Ellison[1]. Some of the most popular

FIGURE 1.1: Types of Popular Online Social Networks

Image courtesy: https://spicensugar.files.wordpress.com/2010/09/social-profit-landscape1.png

OSNs include Facebook, Twitter, YouTube, Quora, LinkedIn, Yelp, Google+ and so on. Fig. 1.1 provides an idea of the number and variety of OSNs that are present. Each of these OSNs has different characteristics and features, and they serve different purposes. For example, in Facebook, the friendship is a bidirectional relationship whereas, on Twitter, follower/following is a unidirectional relationship. In YouTube, the content is video whereas in Yelp the content is mostly textual. LinkedIn is used to connect with

the professional network of people whereas Facebook is used to connect to people one already knows. Quora is used as a Question and Answer platform whereas Twitter is used as a microblogging platform. Thus, it is difficult to generalise the OSNs under one umbrella and yet they have a lot in common. Also, when the popularity of OSNs is on the rise, and the importance of OSNs in the day-to-day life of the layman is increasing, it is important to study and analyse the OSNs.

OSNs have been studied widely over the years. In the initial days, social networks were studied from a mathematical perspective, suggesting generic models for social networks and trying to identify the growth patterns in OSNs[2][3][4]. Mathematically, an OSN is a graph that can be represented using a non-empty set of vertices and a set of edges. The vertices represent entities of the network such as users, posts, videos, tweets and so on and the edges represent the link between these entities such as friends or followers, likes, tags, comments, shares and so on. The vertices and edges might each have different attributes. Earlier studies include the study of random graphs, the small world property, the preferential attachment property and so on.

The study then proceeded to group, clustering and community detection[5][6]. Thus, rather than studying the network as a whole, groups or communities or clusters in the network are selected, and their behaviour as a single entity is observed and analysed. The presence of such groups, the degree of intra-group interactions and so on. are all relevant studies in this era.

Fig. 1.2 provides a gist of the common research topics related to OSNs. In recent

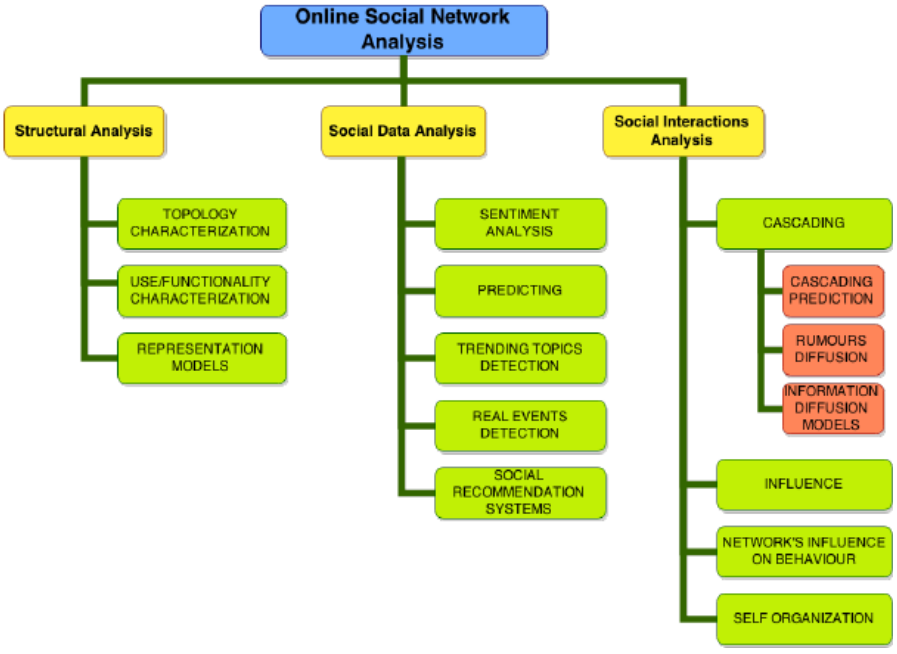


FIGURE 1.2: Research Topics Related to Online Social Networks

Image courtesy: https://www.researchgate.net/profile/Alan_Godoy/publication/275341115/figure/fig1/AS:294465735544833@1447217510420/Categories-of-study-on-Online-Social-Networks-from-a-computational-perspective.png

times, microstructures of the network are being studied[7] and each OSN is treated as a different entity rather than modelling all OSN as large scale networks. Recent studies in OSNs include the behavioural study of the networks, in-depth analysis of network properties and OSN-specific modelling for activity and growth patterns. This

approach finds application mainly in the field of targeted marketing and business modelling. Study of microstructures of OSN combined with behavioural properties can give an insight into individual and group behaviour in the OSN. Attempts have been made

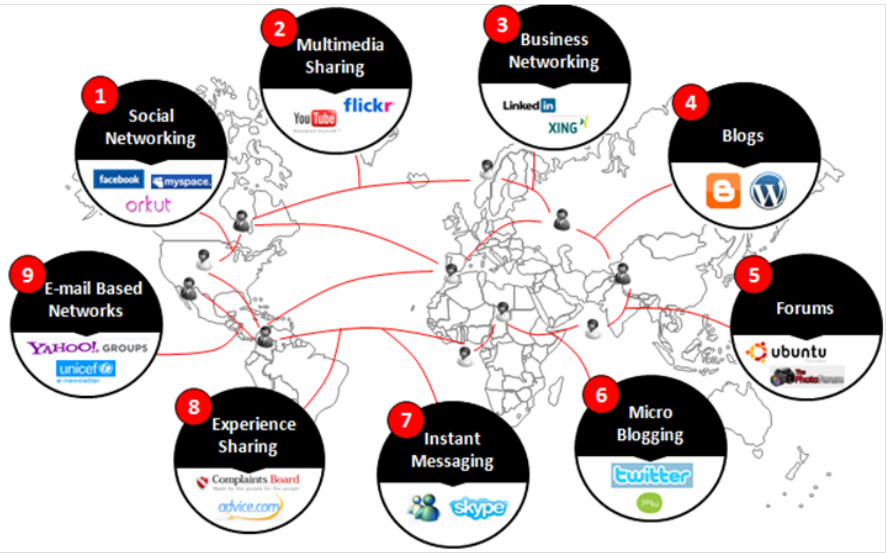


FIGURE 1.3: Popular Categories of OSNs based on purpose

Image courtesy: https://forteconsultancy.files.wordpress.com/2010/02/picture1.png

to categorise OSNs in the past. Fig. 1.3 provides one such categorisation. However, these categorisations do not provide an idea of the network structure and properties of the OSNs.

1.2 Motivation

Currently, although different OSNs are studied individually to understand the microstructures of the network, as a whole, all OSNs as a group are considered to be large scale networks following power law distribution. They are also referred to as scale-free networks[8][9]. While individual networks have been shown to be exceptions to the power law distribution, no prominent literature exists to categorise OSNs based on the network properties. A brief look into the structure of popular OSNs show that although they are apparently different, there are some patterns in the structural properties of the network that can be used to infer the microstructure of the OSN.
Based on the network structure, there also exists several growth models, information diffusion models, rumour blocking models and so on that do not concentrate on the microstructure of the network as individually studying each OSN and devising an OSN-specific model is not feasible. However, the generic models might not be best fits for all OSNs. Thus, studying and categorising the different OSNs based on structural properties can help in feasibly devising category-specific models that are better fits than the generic models.

1.3 Contribution

The purpose of this work is to categorise popular OSNs based on their network structural properties. Initially, a set of relevant network structural properties have been chosen from a massive pool of network properties. These properties are relevant and have significant impact on the structure and microstructure of the network. As a part of this thesis, an extensive study of the network properties of multiple OSNs is carried out. Five real life OSNs have been considered in this research, namely, Facebook, YouTube, Enron Email Network, Twitter, and Google+. All the network properties selected for this thesis are evaluated for each of the five OSNs. The empirical results observed are recorded and represented in tabular format or graphically. The significance of each of the network properties for each of the OSNs have been explained. The empirical results are considered as a basis for categorisation. The effect of the network structural properties on the categorisation has been explained. The OSNs are proposed to be categorised based on the observed similarities and differences in the empirical results. The proposed categories are studied, analysed and justified. Thus, the contributions of this thesis are

1. An extensive empirical study of multiple network properties of five OSNs is carried out.
2. The significance of each of the network properties considered has been explained with the help of the results of the empirical study.
3. The effect of the network properties on the network structure and microstructure has been explained with the help of the results of the empirical study.
4. A categorisation of OSNs based on the network structural properties is proposed.
5. The expected structure and microstructure of each proposed category is explained.

1.4 Organisation

The thesis is divided into six chapters. The first chapter provides an introduction to the topic of the thesis. The second chapter provides a survey of existing state-of-the-art literature in relevant and related topics. The third chapter provides an overview and significance of all the network properties considered for categorisation. The fourth chapter represents the empirical studies and analyses the significance of the observed values. The fifth chapter proposes the categorisation of the OSNs based on the observed values. The sixth and last chapter provides a conclusion and future scope of research in related and relevant topics.

Chapter 2

Literature Survey

2.1 Network Theory Models

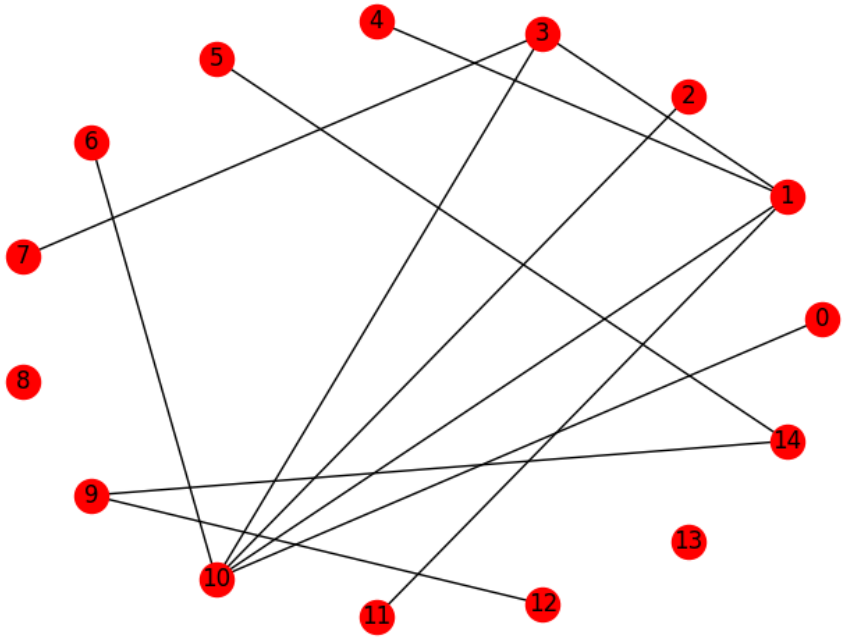


FIGURE 2.1: Random Graph

The study of modern-day networks started with two independent, and yet, quite similar works by contemporary mathematicians. The first and more popular work is by Paul Erdos and Alfred Renyi[2]. In this work, the concept of random graphs was introduced. The concept is that all the graphs that can be drawn from a fixed number of vertices and a fixed number of edges are equally likely. The other work, done by Edgar Gilbert[10], is also about random graphs. This paper says that each edge of a network has a fixed probability of being included in the network. This probability is independent of the other edges in the network. Fig. 2.1 represents a random graph of 15 vertices. Both these papers were published in 1959 and thus began the journey of network analysis. Random graphs follow the Poisson distribution for degree distributions.

The more relevant model for modern-day online social networks is the Barabasi-Albert Model[11] or the Scale-free network model. Fig. 2.2 represents a scale free network of 15 vertices. Scale-free networks are those networks that asymptotically follow power

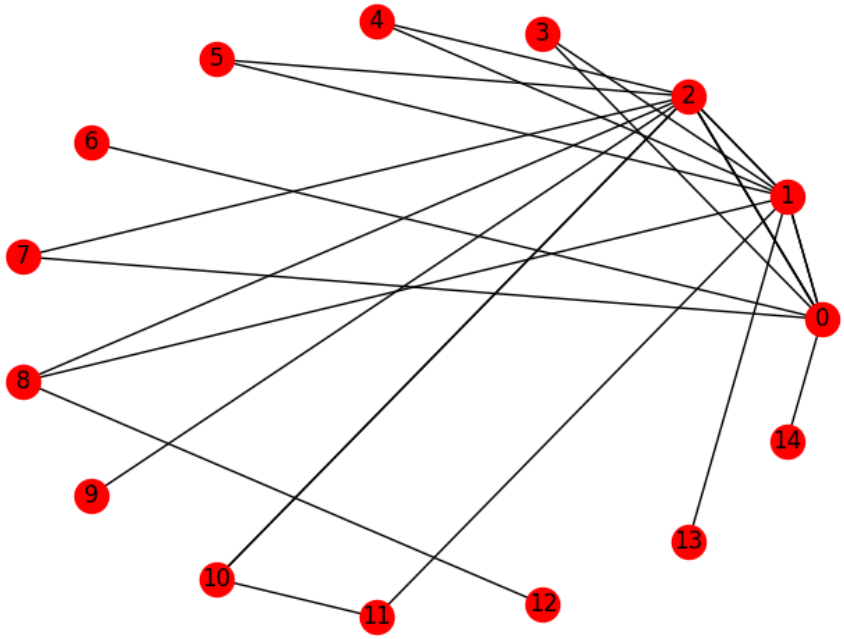


FIGURE 2.2: Scale Free Network

law. Power law can be described using the equation

$$p(k) \propto k^{-\gamma} \tag{2.1}$$

where k is the degree of a node, $p(k)$ is the probability of occurrence of nodes with degree k, and γ is the power law exponent. Thus, in a scale-free network, the probability of high degree nodes is low, and the probability of low degree nodes is high. The power law distribution is a variation of an exponential distribution. All new online social networks are considered to be scale-free networks and are considered to follow power law distributions.

The third kind of network model, also associated with modern day online social networks, is known as the Small World Network. One model for small world networks was identified by Duncan J. Watts and Steven H. Strogatz[4]. It was initially developed as a small world model for random graphs but was later extended to include other network types such as social networks. Fig. 2.3 represents a small world network of 15 vertices. A small world network is one where the average path length of the network grows proportional to the number of nodes in the network, that is,

$$l \propto |V| \tag{2.2}$$

where l is the average path length of the network and V is the vertex set.
l can be defined using the equation:

$$l = \frac{\sum_{\forall i,j \in V} \rho(i,j)}{|E|} \tag{2.3}$$

where $\rho(i,j)$ is the distance between the nodes i and j, V is the set of vertices and E is the set of edges.

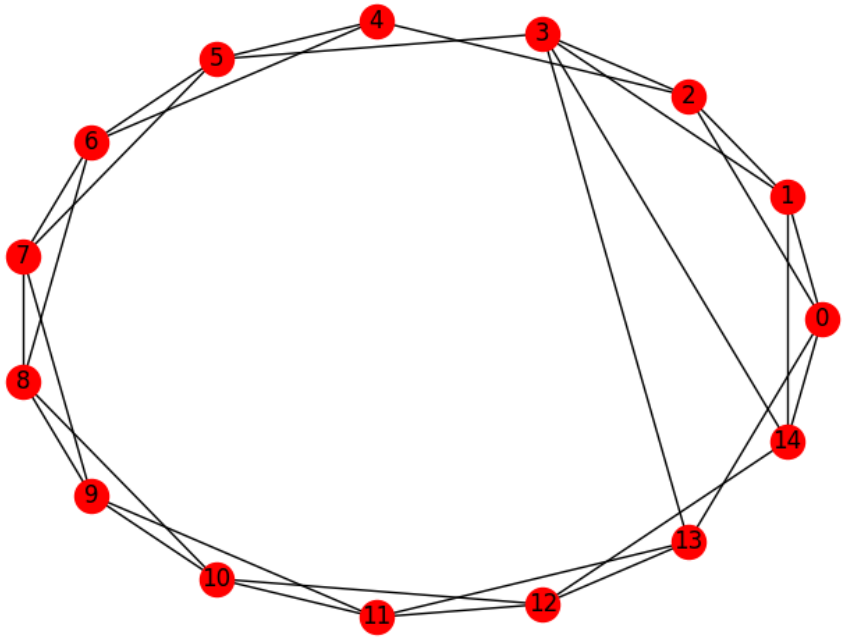


FIGURE 2.3: Small World Network

2.2 Study of Network Properties

Structural properties of the network can be used to identify crucial details about the network structure and microstructures and has multiple applications. Some of the most researched wide domains of applications include information propagation, community detection, study of the growth of the network and so on. Understanding the influence of different structural properties on the dynamics of information flow in an OSN is an interesting open field of research. Different community detection methods utilise different network properties and more approaches can be developed maximising a set of network properties. The effect of network properties on how a network grows is also an interesting problem. Thus, studying the network properties has been a significant field of research.
The most popular group of network properties to be studied are centrality measures. Several centrality measures have been studied for online social networks. Some of the more common centrality measures include - degree centrality, closeness centrality, eigenvector centrality and betweenness centrality.
In 1978, a paper by Linton C. Freeman[12] studied the three centrality measures for online social networks, and that led to a series of developments in measuring an OSN with respect to its centrality measures. In a later paper[13], it is shown that the effect of distance-based centrality is negligible on the structure and behaviour of online social networks. In a paper by Hage and Harary[14], they show that the concept of centre and eccentricity is essential and needs to be included along with the centrality measures for meaningful analysis. Thus, it can be seen that with the study of each new OSN, contradictory and yet, correct, analysis can be made about how an OSN behaves. In [15], the importance of tie strength when analysing different centrality measures, namely, degree, closeness and betweenness centrality, has been shown. This work, thus, considers weighted graphs for centrality measure analysis.
Closeness centrality is a popular centrality measure that has been studied widely in the literature. The work by Linton C. Freeman[12] already establishes closeness centrality

as an important centrality measure. In [16], an efficient algorithm for finding the top-k ranks for closeness centrality for large scale networks has been proposed. Another important centrality measure is eigenvector centrality. This centrality measure has a unique property in contrast to the other centrality measures. Eigenvector centrality can be effectively applied to weighted graphs[17]. The effect of betweenness centrality has also been studied widely in OSNs. In [18], the authors study the betweenness centrality correlation between different categories of OSNs based on assortativity. That is, the authors study the correlation of betweenness centrality in assortative, neutral and disassortative networks. In [19], a new betweenness centrality measure is defined that can be applied to weighted graphs and is dependent on all the independent paths between any two points in the network.
Two other essential properties of OSNs are assortativity and homophily. MEJ Newman in his works [20] and [21] studied the mixing and assortative mixing of nodes in the network. These works study mixing patterns for static characteristics of the nodes such as age, ethnicity and so on. Assortativity can also be studied concerning the similarity in degree centrality. Homophily is another network property that is a measure of similarity. The similarity of nodes can be based on static characteristics such as age and gender, or dynamic characteristics such as friends circle and activity[22][23]. In [24], it has also been shown that networks grow following the principle of homophily.

2.3 Communities, Groups, and Microstructures

The purpose of social network analysis is to get a clear understanding of the OSNs work. One key element of any OSN is the formation of groups or communities in OSNs. A wide field of study has been dedicated to studying the microstructures in OSNs and identifying communities in the OSNs. A study by Leskovec et. al.[25] shows that small and tightly-knit communities, which dissolve at large scale. Another work by Mislove et. al.[9] studies the structure of multiple popular OSNs. The findings of this paper show that these OSNs have an inner structure of connected high degree nodes, and peripheral structures of small tightly-knit clusters of low degree nodes. In [26], it is shown that in most of the OSNs, three general structures exist. The first is isolated nodes, the second is small isolated communities with a star-like topology, and a well-connected central core structure which forms a giant component.
The behaviour of groups or communities over time is another exciting and significant field of study. In [27], the authors study the network structural factors that promote joining new groups and catalyse leaving old groups. The paper also studies the network structural factors responsible for the inclusion of new members in a group. In another paper[28], the authors show the time dependence of community structures in large scale networks. The findings of this paper indicate that the behaviour of small tightly-knit communities and that of the large scale network as a whole are different. In [29], it is shown that in a network consisting of smokers and non-smokers, the size of clusters of smokers remains constant over time but the total number of smokers decrease with time. Thus, smokers quit in clusters, which is another example of group dynamics.

2.4 Social Network Analysis in Marketing

Currently, the main application of social network analysis is in marketing and advertising. In [30], the presence of brand communities in present-day OSNs, and how it can be beneficially exploited has been studied. In [31], the effect of different network properties on marketing studies. The paper shows the how different structural properties can be used to better understand relationships amongst members of a business-to-business network. In [32], the effect of network properties on research in the field of hospitality management and marketing is highlighted. Studies in this paper show that most academic collaborators are located in close geographic proximity in a network of co-authors.
The future of online social network analysis in the field of marketing has been studied in [33] amongst recent studies in this domain. Another recent work which comes as an application of online social network analysis in marketing includes the work in [34]. In this work, green consumption related published articles were considered and community detection was applied in order to find the main theoretical relationships.

2.5 Previous Categorisation of OSNs

The categorisation of OSNs can be interpreted in multiple ways. Most attempts of categorisation or classification on OSNs are related to node classification, relationship classification or classification of the content present on online social networks. However, little literature exists on categorising OSNs based on the structural network properties. Node classification attempts such as the network embedding approach preserving the structural as well as attribute proximity[35] or classification of nodes in streaming data, i.e., the massive volume of data generated continuously[36]. These works concentrate on classifying nodes based on some learning or training and do not focus on categorising the OSN as a whole. Another form of classification on OSN includes relationship classification. In [37], the authors first study and categorise the different types of relationships in the network based on their network properties as well as behavioural properties. They then exploit this classification to optimise information propagation in OSNs. Content classification is another well-researched area in OSNs. In [38], the authors present a feature selection technique for large scale short textual data for real-time classification. The feature selection is based on both the social network data and the content-based data and thus provides more accurate results.

2.6 Summary

It can be seen by reviewing the existing literature, that the network properties have been extensively studied in the past and their results thoroughly analysed. It can also be seen that the network microstructures have also been studied thoroughly and efficient methods have been suggested for identifying and extracting the microstructures. Literature also exists on the classification of nodes, edges and content based on both the network structural properties as well as the attributes and behavioural properties.
However, examples of combining multiple network properties to categorise the OSNs

as a whole, and not their nodes, edges or content individually, can be found in the literature. Also, no study of structure and microstructures of an OSN based on categorisation is present in literature. This kind of an empirical study can be crucial to quickly identify, analyse and process OSNs, especially because of the growth of existing networks and the inclusion of new OSNs. This research work, thus, provides a unique categorisation of OSNs based on structural properties and the correlation of the categories with network structure and microstructures. The next chapter provides a theoretical study of different network properties and structures that are relevant for understanding the structure and microstructures of an OSN.

Chapter 3

Network Structural Properties

3.1 Degree Centrality

One of the most widely studied network property is Degree Centrality. Degree centrality can be defined as *the number of edges connected to a node*. For directed networks, degree centrality can be of two types - out-degree and in-degree. Out-degree is the number of edges originating from a node and in-degree is the number of edges ending at a node.[39]
Degree centrality is a measure of connectedness or importance of a node. The more the degree of a node (i.e., the more the number of edges connected to a node), the more is it is "importance". The "importance" can be studied from many perspectives. From the network structure perspective, high degree centrality nodes contribute more towards network growth following the preferential attachment rule.
The preferential attachment rule follows the *rich gets richer* phenomena. The rule is that higher the degree of a node more is its probability of attaining new edges. The new edge might be between two existing nodes or between a new node and an existing node. By principle, it is more probable for a new node to attach to an "important" node, thus making it more "important". However, the preferential attachment rule does not consider the activity level or inherent bias of a node.
A node can have a high degree centrality but might be inactive in the network, which can decrease it is "importance". Moreover, a node might be inherently "popular" outside the network which might affect the popularity of the node in the network. These two cases can lead to an exception to the preferential attachment rule.
Most large scale networks follows Power Law Distribution[40][11]. The preferential attachment rule leads to the power law distribution of online social networks. A power law distribution has a probability distribution function

$$p(x) = \frac{\alpha - 1}{x_{min}} \frac{e^{(ln\, x - \mu)^2 / 2\sigma^2}}{x\sigma\sqrt{2\pi}} \tag{3.1}$$

Degree centrality is generally represented using a degree distribution plot. The degree distribution can be represented using a pdf plot, a log-log plot of the pdf and cumulative degree distribution. The distributions can be represented using both uniform and non-uniform bin sizes. The power law distribution has a signature curve in the pdf and CDF plot and is a straight line with a negative slope in the log-log plot.

3.2 Power Law Exponent

Power law exponent is *closely connected to the degree distribution of the network*. Power law exponent *describes the shape of the long-tail distribution*. The study of networks started with random networks[2]. The degree distribution for random networks follows the Poisson distribution. However, most real-life large scale networks, whether it be the World Wide Web[3] or Co-author network[40] so on. all follow Power law distribution[41]. The power law distribution can be calculated as follows -

$$p(k) \propto k^{-\gamma} \tag{3.2}$$

where γ is the power law coefficient. Thus, the number of nodes with degree k decreases exponentially with a factor of γ as the value of k increases. In other words, high degree nodes are infrequent, and low degree nodes are frequent.
Given the degree distribution of a network, the power law exponent can be estimated as follows -

$$\gamma = 1 + n \left(\sum_{u \in V} ln \frac{d(u)}{d_{min}} \right)^{-1} \tag{3.3}$$

where, V is the vertex set, $d(u)$ is the degree of node u and d_{min} is the minimum degree threshold.

3.3 Connected Components

Number of components is *a measure for the group formation and connectivity of the network*. A connected component in a network is a sub-graph in which there exists a path between every pair of vertices $i, j \in V'$ where V' is the set of vertices of the sub-graph. For directed networks, there are two types of connected components - weakly connected components and strongly connected components. A weakly connected component is a sub-graph in which if the direction of the edges is disregarded, it becomes a connected component. A strongly connected component is a sub-graph in which there exists a directed path between every pair of vertices $i, j \in V'$ where V' is the set of vertices of the sub-graph. A network is said to be connected if the number of connected components(strongly connected network for directed networks) is 1 and disconnected otherwise.

3.4 Path Lengths

The diameter of a network gives a sense of the *vastness or scale of the network*. The distance($\rho(i, j)$) between two nodes $i, j \in V$ is the minimum number of edges between the two nodes, where V is the vertex set. The diameter of the network is given as -

$$max(\rho(i, j)) \forall i, j \in V \tag{3.4}$$

That is, the diameter of a network is the maximum distance of all calculated distances in the network[42].

Average Path Length of a network is another measure of the *vastness or scale of the network*. It is given by the formula -

$$\frac{\sum_{i,j \in V} \rho(i,j)}{|V|^2} \tag{3.5}$$

That is, the average path length of a network is the average of all calculated distances in the network.
The OSNs are huge if measured in terms of the number of nodes or number of edges. Therefore, they can be expected to have large values for diameter and average path lengths. However, all OSNs have small diameters and average path lengths below 6[43].
If a network is disconnected, then path lengths such as diameter and average path length are calculated for the biggest connected component(that is, the connected component with the highest number of nodes).

3.5 Density

Density is *the ratio of the number of edges present in the network to the total number of edges possible in the network*[42]. The maximum possible number of edges in a directed network can be -

$$|V|(|V|-1) \tag{3.6}$$

and the maximum possible number of edges for the undirected network can be -

$$\frac{|V|(|V|-1)}{2} \tag{3.7}$$

where, $|V|$ is the number of nodes in the network.
Therefore, for directed network the value for density of the network is calculated as -

$$\frac{|E|}{|V|(|V|-1)} \tag{3.8}$$

and that for undirected network as -

$$\frac{2|E|}{|V|(|V|-1)} \tag{3.9}$$

where, $|E|$ is the number of edges in the network.

3.6 Clustering Coefficient

Clustering Coefficient is a measure of *denseness of a network and its vertices*. In other words, the clustering coefficient measures the denseness of the neighbourhood of a node. There are three types of clustering coefficients -

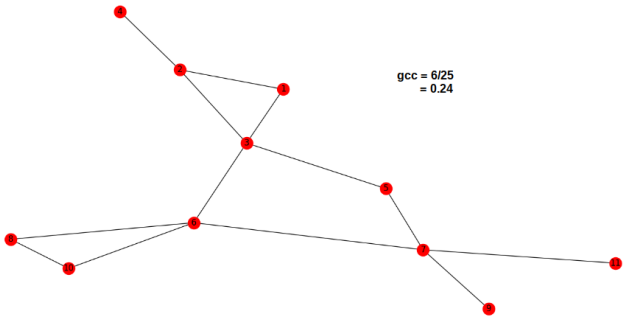


FIGURE 3.1: Global Clustering Coefficient

3.6.1 Global Clustering Coefficient

Global Clustering Coefficient is a denseness measure for the network as a whole[42]. It is calculated as -

$$\frac{3 * number\ of\ triangles}{number\ of\ triplets} \tag{3.10}$$

where, a *triplet* is a connected sub-graph of 3 vertices and a *triangle* is a cycle of size 3. Fig. 3.1 represents the global clustering coefficient of a network.
Higher is the value of the global clustering coefficient; denser is the network. However, the global clustering coefficient by itself is not a sufficient measure for network denseness. This is because, parts of the network might be very dense while other parts might be sparse, but the global clustering coefficient might be dense which can be misleading.

3.6.2 Local Clustering Coefficient

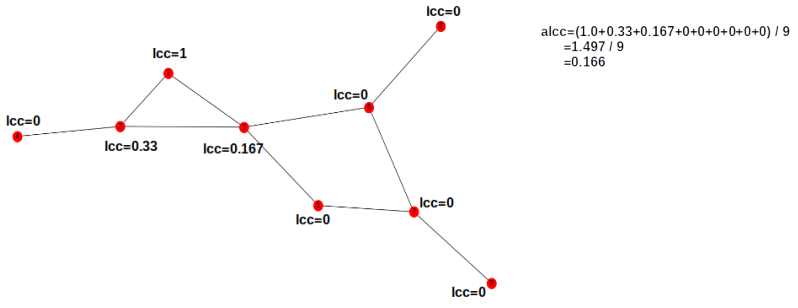


FIGURE 3.2: Local Clustering Coefficient and Average Local Clustering Coefficient

Local Clustering Coefficient is a denseness measure for the neighbourhood of a node. For directed networks, the local clustering coefficient is calculated as -

$$\frac{e_i}{k_i(k_i-1)} \tag{3.11}$$

and that for undirected network as -

$$\frac{2e_i}{k_i(k_i-1)} \tag{3.12}$$

where k is the number of neighbours of a node and e is the number of edges present in the network between the k neighbours.
The local clustering coefficient for every node of a network might be represented as a distribution curve or as histograms.
The average local clustering coefficient for the whole network can be calculated as follows -

$$\frac{\sum lcc(i)_{i\in V}}{|V|} \tag{3.13}$$

where V is the set of vertices and $lcc(i)$ is the local clustering coefficient of node i and $|V|$ is the number of nodes in the network.
Fig. 3.2 depicts the local clustering coefficient and the average local clustering coefficient of a network. The average local clustering coefficient is a measure of average denseness of the network[4]. However, average local clustering coefficient by itself is not a sufficient measure for network denseness. This is because, parts of the network might be very dense while other parts might be sparse, but the average global clustering coefficient might be dense which can be misleading.
Thus, to have a clear understanding of the network structure, all the different types of clustering coefficient needs to be considered.

3.7 Cliques and Hubs

3.7.1 Cliques

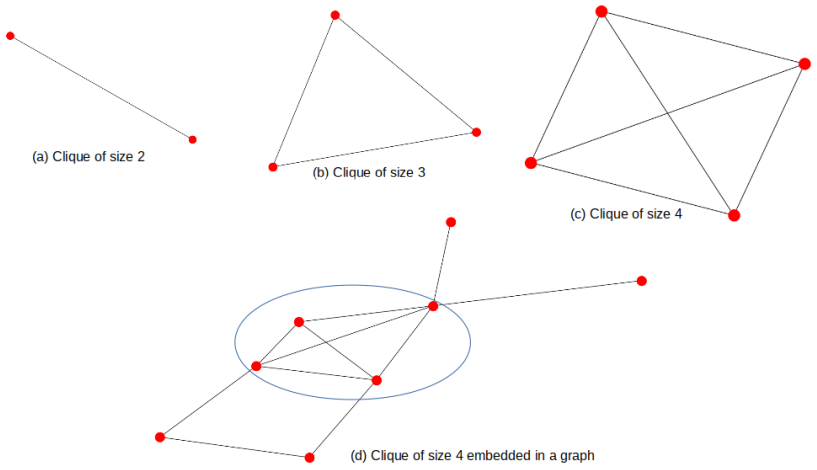


FIGURE 3.3: Cliques

Cliques are complete sub-graphs. They are the densest sub-graph possible, where every node in the sub-graph is adjacent to all other nodes in the sub-graph. Fig. 3.3 represents complete subgraphs of size 2, 3, and 4 and a clique of size 4 embedded in a graph. Thus, the density, average local clustering coefficient and global clustering coefficient of such a sub-graph are 1, the local clustering coefficient of every node is also 1. However, finding such sub-graphs in an OSN is unlikely. A pseudo-clique is a relaxed form of a clique where some of the edges of the clique might be missing. There is no standard threshold percentage of edges that needs to be present in a sub-graph for it to be a pseudo-clique.

3.7.2 Hubs

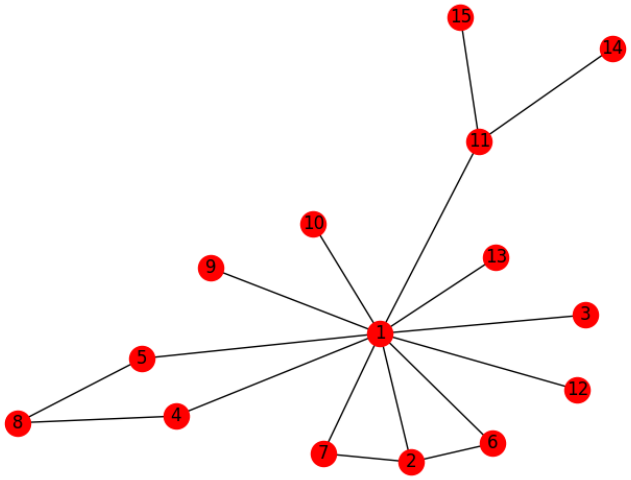


FIGURE 3.4: Hub

Star is a sub-graph where one node is connected to all other nodes in the sub-graph. The central node which is adjacent to all other nodes in the sub-graph is called a Hub. All nodes in the sub-graph, other than the hub, are not adjacent to each other. However, finding stars in OSNs is unlikely. Star-like structures are more common in OSNs, where some of the nodes, other than the hub, in the sub-graph might be adjacent. There is no standard upper bound on how many such edges can be present before a sub-graph can no longer be called a star-like structure.

3.8 Closeness Centrality

Closeness centrality is a *relative closeness of nodes in a network*. It is the reciprocal of distances of all nodes from a particular node i in the network[44] and it can be calculated as -

$$c(i) = \frac{1}{\sum_{\forall j} \rho(i,j)} \tag{3.14}$$

where node j and node i are in the same connected component. High closeness centrality means the node is closely connected to all nodes in the connected component.

3.9 Eigenvector Centrality

Eigenvector centrality is a *measure of the importance of the neighbourhood of a node*[4]. Eigenvector centrality can be measured as -

$$e(i) = \frac{1}{\lambda} \sum_{\forall j} g_{ij} e(j) \tag{3.15}$$

where i and j have a path between them, λ is a proportanality constant, g_{ij} is 1 if there exists an edge between i and j, 0 otherwise. High eigenvector centrality means the neighbourhood of a network is important.

3.10 Betweenness Centrality

Betweenness centrality is a *measure of the connectedness of a network*. Betweenness centrality considers the number of paths that passes through a node i. These paths may or may not originate or terminate at node i. The betweenness centrality can be calculated as -

$$b(k) = \frac{2}{(|V|-1)(|V|-2)} \sum_{i,j,k \neq i,k \neq j} \frac{\nu(i,j,k)}{\nu(i,j)} \tag{3.16}$$

for undirected graphs and

$$b(k) = \frac{1}{(|V|-1)(|V|-2)} \sum_{i,j,k \neq i,k \neq j} \frac{\nu(i,j,k)}{\nu(i,j)} \tag{3.17}$$

for directed graphs. Here $\nu(i,j)$ is the number of shortest paths between node i and j, and $\nu(i,j,k)$ is the number of shortest paths between node i and j that passes through node k.
Betweenness centrality measures how important a node is in connecting other nodes. So, if a node with high betweenness centrality is removed from a network, the network would be a lot less connected.

3.11 Homophily

Homophily is the measure of *similarity of nodes*. A vertex in a graph can have multiple attributes. Some of the attributes can be static, such as demographic data, while other attributes can be dynamic, like friend's list, activity level and so on. Edges of a graph can also have multiple attributes. The similarity of vertices can be measured based on vertex attributes as well as edge attributes. However, not all attributes of a vertex or a node can be considered as structural property. Rather, these attributes can be termed as *beavioral properties*.
Moreover, it is difficult to obtain vertex and edge attributes for OSNs due to security issues and privacy policies adopted by the OSNs.
For this thesis, only one structural property, namely, the clustering coefficient has been considered for judging the similarity of two nodes. If two nodes have similar local clustering coefficients, then they are described to be similar. The vertices are divided into clusters or groups based on their local clustering coefficient values, where nodes

with similar local clustering coefficient are grouped. An algorithm proposed by Sen et. al.[45] has been used or this clustering.

3.11.1 Clustering Algorithm

For this thesis, a clustering algorithm by Sen et. al.[45] has been considered. This algorithm divides vertices of a network into non-empty structures referred to as focal structures. The algorithm calculates the local clustering coefficient $lcc(i), \forall i \in V$ where V is the set of vertices. The algorithm also calculates the average local clustering coefficient $alcc$ for the network. Then, for every pair of vertex $i, j \in V$ the algorithm calculates $lcc(i)$ and $lcc(j)$. If both $lcc(i) > alcc$ and $lcc(j) > alcc$ or both $lcc(i) < alcc$ and $lcc(j) < alcc$ the the vertex i, j are allocated to the same focal structure, say f_i. This algorithm is appropriate because it takes into account only two simple network property metrics, namely, local clustering coefficient and average local clustering coefficient, and also has a low computational time complexity which is desirable for working with large scale networks.

After clustering, the homophily value of the network is calculated as[42] -

$$h_s^t = \begin{cases} 1, if s = t \\ 0, otherwise \end{cases} \tag{3.18}$$

where, s and t are two groups as divided by the algorithm proposed by Sen et. al.[45]. This value is calculated for all pairs of s and t. A value of 1 means all edges within the group and a value of 0 means all edges are inter-group edges[24].
The average homophily value of the whole network can be calculated as -

$$h = \frac{h_s^t}{c} \tag{3.19}$$

where c is the number of h_s^t with non-zero values. Average homophily value for a directed network can be calculated as h_{in} and h_{out} for in-degree network and out-degree network respectively.
Higher values of average homophily for the network signifies that similar nodes are well-connected in the network and lower homophily values indicate that similar nodes are disconnected in the network.

3.12 Assortativity

Assortativity is a measure of *connectedness of similar nodes*. Here, the similarity is implicitly considered in terms of degree centrality. Assortativity can assume positive values and negative values. The range of assortativity values lie within $[-1, 1]$
The assortativity value for a network can be calculated as[20][21] -

$$r = \frac{\sum_{jk} e_{jk} - q_j q_k}{\sigma_q^2} \tag{3.20}$$

where

$$q_k = \frac{(k+1)q_{k+1}}{\sum_j jp_j} \tag{3.21}$$

and

$$\sigma^2 = \sum_k k^2 q_k - \left(\sum_k k q_k\right)^2 \tag{3.22}$$

The value of r can be interpreted as -

$$r \begin{cases} \approx 0, neutral \\ > 0, assortative \\ < 0, disassortative \end{cases} \tag{3.23}$$

where, in assortative networks, nodes with a similar degree are connected, and nodes with a dissimilar degree are disconnected. In disassortative networks, dissimilar nodes are connected, and similar nodes are disconnected. In neutral networks, either both similar as well as dissimilar nodes are disconnected, or both are connected. Assortativity value for a directed network can be calculated as r_{in} and r_{out} for in-degree network and out-degree network respectively.

3.13 Summary

In this chapter, the network structural properties have been studied in details. These properties form the basis of the empirical study that has been conducted in the next chapter. Five real-life OSNs have been considered for empirical study of the network structural properties. Each of the property is empirically studied for each of the OSN. The network properties studied in this chapter, either individually or collectively, help in inferring some structural or microstructural details of the OSNs. The inferences are drawn based on the results of the empirical study that has been conducted in the mext chapter.

Chapter 4

Empirical Study of Network Structural Properties

4.1 Overview

The previous chapter outlines the major network structural properties. These properties have been empiricaly studied in this chapter. The network structural properties have an expected value or expected range of values for the OSNs. The observed values for the network structural properties have been compared with the expected values or range of values. The similarities and differences in the observed and expected values are acknowledged, analysed and explained to better understand the network structure and microstructure of the OSNs.

4.2 Selection of Networks

The first step in conducting this research was to identify and select OSNs. In current times, there is a wide variety of OSNs available. However, considering the complete set of networks is not feasible for this study. Therefore, the challenge was to select a sample that would provide a correct and accurate representation.
For this research, five online social networks were chosen, namely, Facebook, YouTube, Email, Twitter, and Google+. This set provides much variety in terms of purpose, type of relationship, nature of the relationship, network density and many such other network metrics.

4.2.1 Facebook

Facebook was founded in 2004 in the United States of America and made worldwide debut in 2005. It is listed as a social networking site. Currently, there are more than 2.38 billion active user profiles on Facebook. The nature of the relationship of Facebook is friendship, which is a bidirectional relationship(unless explicitly restricted by the user). The nature of the content on Facebook can be textual, audio-visual and other multimedia types. Facebook was primarily used to connect, share and stay in touch with people one already knows. Facebook has two other popular OSNs, namely, Instagram and WhatsApp, as its subsidiaries. The dataset has been collected from http://konect.uni-koblenz.de/networks/facebook-wosn-links. It is an undirected

network of friends relationship. The number of nodes in the network is 63,731 and the number of edges is 8,17,035.

4.2.2 YouTube

YouTube was founded and launched in 2005 in the United States of America and has been active worldwide ever since. It is listed as a Video hosting site. YouTube currently works as a subsidiary of its parent company Google. The type of content on YouTube is exclusively audio-visual. However, follow-up communication based on the audio-visual content is textual. Although holding a profile on YouTube is not compulsory for viewing contents, it is mandatory for any form of communication. The nature of the relationship on YouTube is friendship, which is bidirectional. Currently, YouTube has 1.3 billion user profiles. The dataset has been collected from http://socialcomputing.asu.edu/datasets/YouTube. In this dataset, 5 edge sets are provided. Out of the five edge sets, "1-edges.csv" was considered. It is an undirected network of friends relationship. The number of nodes in the network is 13,723 and the number of edges is 76,765.

4.2.3 Email

The email network considered for this research is known as the Enron Email network. It is a set of emails that were sent/received by the employees of the American company Enron. The data was initially made public by the Federal Energy Regulatory Commission. There were multiple integrity issues in the initial dataset, and the version of the dataset that has been used for this research has all the integrity issues resolved. The dataset contains both professional as well as personal emails. The emails sent/received by non-employees of Enron has also been considered in the dataset, as long as at least one of the sender or receiver of the email is an Enron employee. Email-based networks are also considered as online social networks, and thus this network has been included for this study. The nature of the relationship for this network is email. The dataset has been collected from https://snap.stanford.edu/data/email-Enron.html. In this edge set, each undirected edge has been represented using two directed edges. Thus, it is considered as an undirected network of email contacts. The number of nodes in the network is 36,692 and the number of edges is 1,83,831.

4.2.4 Twitter

Twitter was founded in 2006 in the United States of America and has been active worldwide since then. It is listed as a news and social networking site. There are currently more than 300 million active users on the network. Twitter supports multimedia and textual content and is often referred to as a microblogging site. Initially, there was a 140 character restriction on the published text, and only textual content was supported, following the Short Message Service(SMS) ideology. Currently, the limit on textual content is 280 characters. The type of relationship on Twitter is unidirectional. Contrary to Facebook, Twitter is primarily used to connect to popular figures. Currently, Twitter has three other OSN subsidiaries, namely, Vine, Periscope and MoPub.The dataset has been collected from http://konect.uni-koblenz.de/networks/munmun_twitter_social. It

is a directed network of follow contacts. The number of nodes in the network is 4,65,017 and the number of edges is 8,34,797.

4.2.5 Google+

Google+ was founded in 2011 by Google in an attempt to replace a prior Google product Google Buzz. Google+ was shut down in April 2019. It was listed as a social networking service. In 2015, Google+ had 111 million active user profiles. The nature of the content on Google+ was multimedia and textual, and the type of relationship on Google+ was friends circle, which is unidirectional. Google+ was primarily launched as a challenge to contemporary OSNs and was connected to other Google products like Bloggr, YouTube and Gmail. Google+ has an Enterprise version which is still in use by G-Suite users for intra-organisation communication. The dataset has been collected from http://konect.uni-koblenz.de/networks/ego-gplus. It is a directed network of friends circle. It is an ego network, where there are central *hubs* or ego nodes and peripheral nodes or alter-egos. The number of nodes in the network is 23,628 and the number of edges is 39242.

It can be seen from the network descriptions provided above that none of the networks considered for this research is a complete representation of the OSN, but rather, is a small sample of the actual network. Thus, some errors might be present in the observed values in following sections due to the sampling errors.

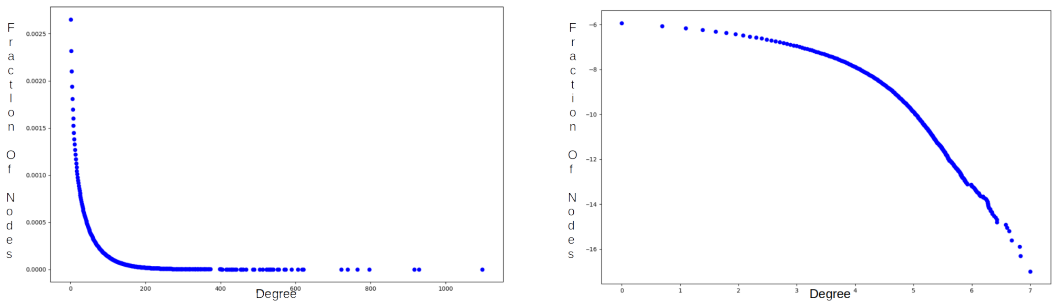
4.3 Results and Analysis

4.3.1 Degree Distribution

Degree distribution of a network is the probability of a node to have degree k. It is generally represented graphically. For this research, the cumulative density function for the degree distribution has been considered for plotting. The cumulative density function for degree distribution can be represented as -

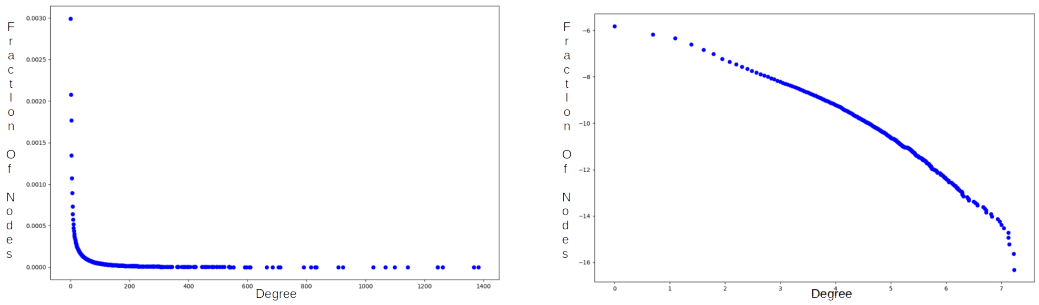
$$p(i) = \frac{count(j|j \geq i)}{|V|} \quad \forall i \in 1 \rightarrow d_{max} \tag{4.1}$$

where $count(j|j >= i)$ is the number of degree values greater than or equal to i, V is the set of vertices, and d_{max} is the maximum degree in the network.
For this thesis, the cumulative density function for the degree distribution has been represented in two ways. The normal distribution has been represented with the log-log plot represented inset. The graphical representations contain degree on the x-axis as the independent variable and the fraction of nodes with a particular degree on the y-axis as the dependent variable. Fig. 4.1 represents the degree distribution for the Facebook network. It can be observed that both the normal plot and the log-log plot of degree distribution follows the shape of power law distribution. Fig. 4.2 represents the degree distribution for the Email network. It can be observed that both the normal plot and the log-log plot of degree distribution follows the shape of power law distribution. Fig. 4.3 represents the degree distribution for the YouTube network. It can be observed that both the normal plot and the log-log plot of degree distribution follows the shape of



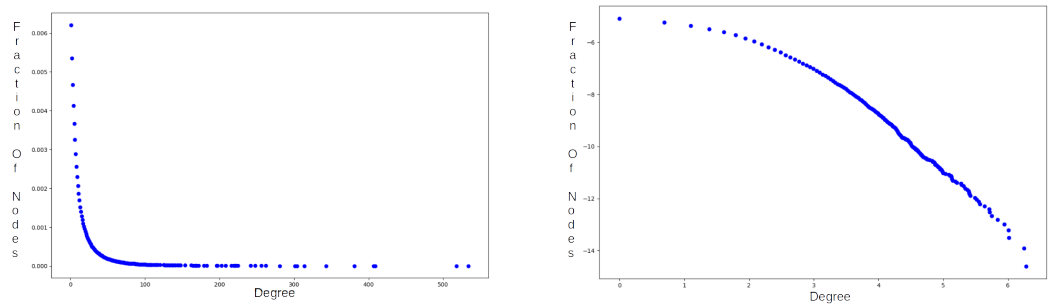
(A) Cumulative Density Distribution (B) Log-Log Plot

FIGURE 4.1: Facebook Degree Distribution



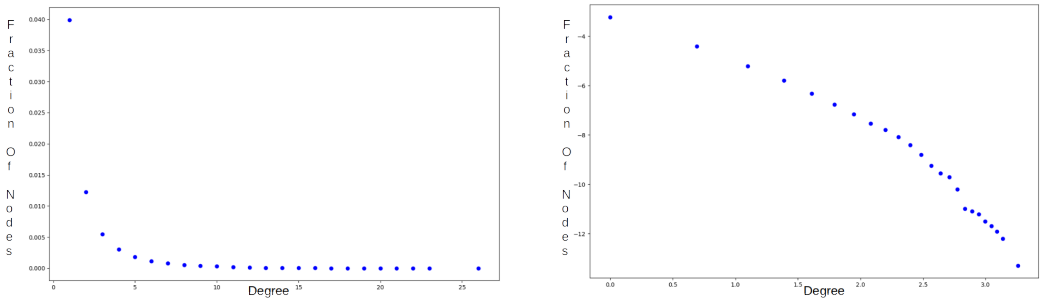
(A) Cumulative Density Distribution (B) Log-Log Plot

FIGURE 4.2: Email Degree Distribution



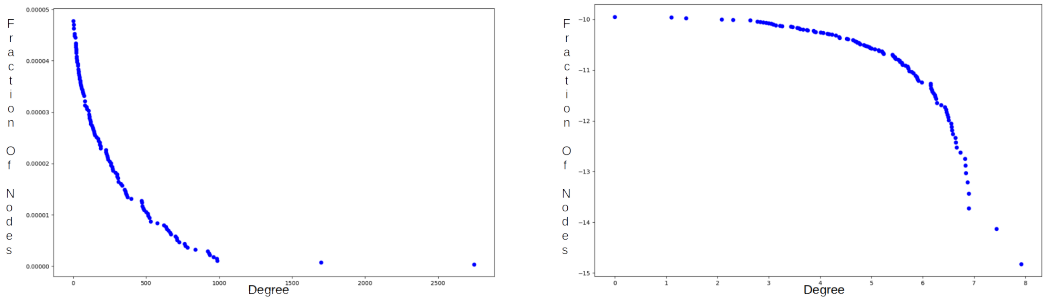
(A) Cumulative Density Distribution (B) Log-Log Plot

FIGURE 4.3: YouTube Degree Distribution



(A) Cumulative Density Distribution (B) Log-Log Plot

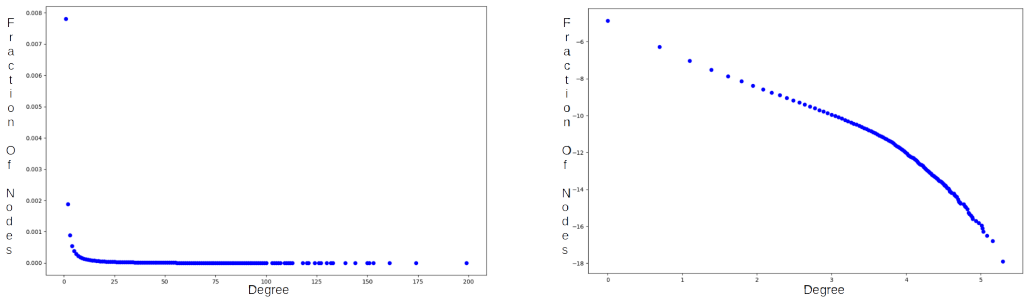
FIGURE 4.4: Google+ In-Degree Distribution



(A) Cumulative Density Distribution (B) Log-Log Plot

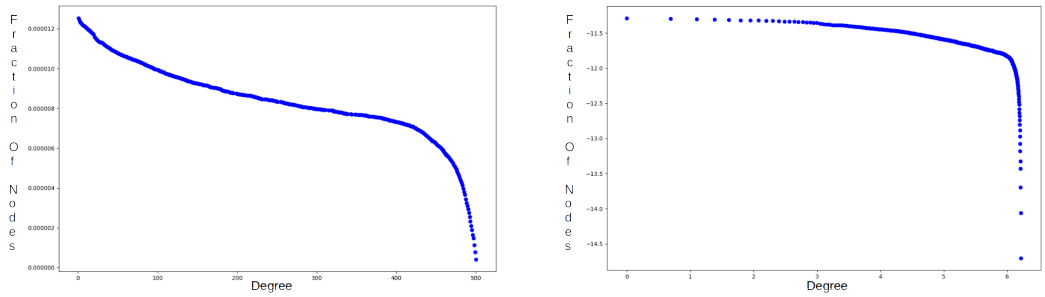
FIGURE 4.5: Google+ Out-Degree Distribution

power law distribution.
It can be seen that the normal plots and the log-log plots of all three undirected networks are close to the cumulative density function of power law. However, there are small deviations that can be observed from these figures. These deviations suggest that the undirected networks, although are similar to power law distributions, do not strictly follow power law distributions. For directed networks, there are two degree distributions, in-degree distribution and out-degree distribution. Fig. 4.4 and Fig. 4.5 represents the in-degree and out-degree distribution of the Google+ network respectively. Similarly, Fig. 4.6 and Fig. 4.7 represents the in-degree and out-degree distribution of the Twitter Network respectively. It can be seen that although the in-degree distributions are somewhat similar to the power law distribution, the out-degree distributions of these networks do not follow power law. The stark difference in in-degree



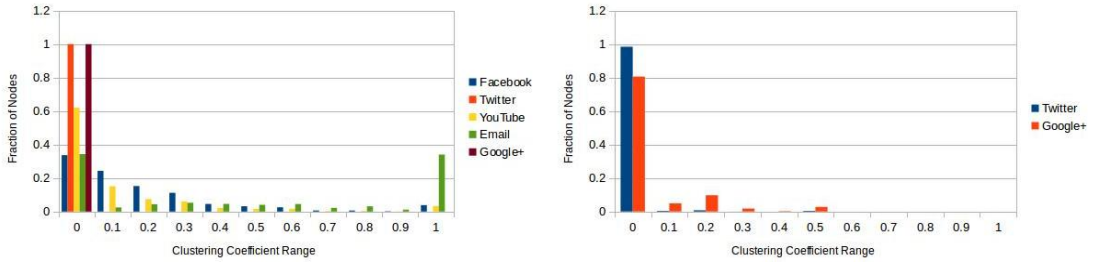
(A) Cumulative Degree Distribution (B) Log-Log Plot

FIGURE 4.6: Twitter In-Degree Distribution



(A) Cumulative Degree Distribution (B) Log-Log Plot

FIGURE 4.7: Twitter Out-Degree Distribution



(A) Local Clustering Coefficient Distribution of OSNs(out-degree)

(B) Local Clustering Coefficient Distribution of OSNs(in-degree)

FIGURE 4.8: Local Clustering Coefficient Distribution

and out-degree networks for directed networks show that even the same network can have very different structure and behaviour if the direction of the edges are changed. Observing the degree distributions of undirected networks, and the in-degree distributions of the directed networks closely show that they can be represented as a composite distribution. Visually observing the shape of the curves show that power law and log-normal distributions are the two strongest candidates. A combination of these two distributions with power law fitted to the tails of the distribution and lognormal fitted to the body of the distribution can provide the best fit.

4.3.2 Local Clustering Coefficient Distribution

Fig. 4.8a and Fig. 4.8b represents the local clustering coefficient distribution of the five OSNs that have been considered for this research. Since changing the direction of edges in directed networks yield different neighbourhoods, the local clustering coefficient distribution for the in-degree networks have been represented in Fig. 4.8b. It can be seen from Fig. 4.8a that in both the directed networks, almost all the nodes have local clustering coefficient in the range of 0 to 0.1. However, for the undirected networks, all ranges of local clustering coefficients have some nodes. The higher number of nodes have low local clustering coefficient, and lower number of nodes have high local clustering coefficient. However, for the case of the email network, there is an unlikely spike between the range of 0.9 to 1 of the local clustering coefficient. This indicates that either the email network has a lot of cliques or pseudo-cliques, or a high number of isolated connected pairs of nodes.

4.3.3 Structural Properties

There are multiple network structural properties that have been studied in literature. An extensive set of structural properties have been selected for the purpose of this paper. The selected properties encompass all the important features of an OSN. The structural properties considered in this section for the empirical study includes -

1. Power Law Exponent
2. Connected Components
3. Diameter
4. Average Path Length
5. Density
6. Global Clustering Coefficient
7. Local Clustering Coefficient
8. Average Local Clustering Coefficient
9. Closeness Centrality
10. Betweenness Centrality
11. Eigenvector Centrality
12. Homophily
13. Assortativity

Power Law Exponent

The value of power law exponent(γ) generally lies in the range of 2 to 3 for OSNs. The power law exponent values observed for the five OSNs under consideration show a particular trend. The first thing to note that, since N_4 and N_5 are directed networks, they can have two different degree distributions, one for the in-degree and the other for the out-degree. Thus, for N_4 and N_5, the power law exponent for in-degree can be presented as γ_{in} and that for out-degree distribution can be presented as γ_{out}.
Empirical study shows that the values of γ for all the networks are within the range of 2 to 3, except the out-degree distribution of N_4 and both in-degree and out-degree distribution of N_5. Facebook, Email and YouTube have power law exponent values of 2.4, 2.4, and 2.1 respectively. For Twitter and Google+, power law exponents are given as $\gamma_{out} = 2.9$, $\gamma_{in} = 4.7$ and $\gamma_{out} = 1.5$, $\gamma_{in} = 4.2$ respectively. Thus, it can be assumed that the OSNs follow power law with few exceptions Another important point to note is the value of d_{min}, which is not always 0. The significance of d_{min} is that the distributions follow the power law with the give value of γ for $d(i) \geq d_{min}$ where $d(i)$ is the degree of the i^{th} node. Thus, if d_{min} is not equal to 0 or 1, that implies that parts of the network may not strictly follow power law.
Thus, based on the values of γ and d_{min}, it can be concluded that the networks do not strictly follow power law throughout the distribution.

Connected Components

For directed networks, both the number of weakly connected components as well as strongly connected components has to be considered. It can be seen from the results of the empirical study that the number of strongly connected components of directed networks is almost the same as the number of nodes in the network. However, the number of weakly connected components is less in directed networks. The number of strongly connected and weakly connected components for Twitter are 1 and 4,63,245 and that for Google+ are 4 and 23,571 respectively. In comparison, the number of connected components for undirected networks is much lower than the number of strongly connected components but higher than the number of weakly connected components. The number of connected components for Facebook, Email and YouTube are 144, 21, and 1065 respectively. Thus, it can be said that undirected networks have more well-formed groups as compared to directed networks if each connected component is considered as a group.

Diameter

The diameter of an OSN does not grow proportionally with the number of edges or number of nodes in a network. In general, the diameter of a network has a small upper bound. For disconnected networks, the diameter of the largest connected component is considered as the diameter of the network. Multiple values have been stated for diameter as the values presented in http://konect.uni-koblenz.de/networks did not always match with the observed value. Thus, both values have been presented for a complete picture. The diameter of Facebook, YouTube, Email, Twitter, and Google+ as calculated for the empirical study are 15, 12, 13, 8, and 8 respectively. However, the values for diameter of Twitter and Google+ is 19 and 10 respectively as shown in http://konect.uni-koblenz.de/networks. Thus, it can be seen from the empirical study, that the value of diameter for both directed as well as undirected network is below 20 in every case, irrespective of the size of the network.

Average Path Length

The value of average path length generally lies between the range of 4 to 6 for OSNs. Thus, the average path length does not grow proportionally with the number of nodes or number of edges. The average path lengths for the five OSNs, as observed from the empirical study, for both the directed as well as undirected networks is below 6, as predicted. The average path length for Facebook, YouTube, Email, Twitter, and Google+ are 4.28, 4.26, 3.39, 4.59, and 3.95 respectively.

Density

Online social networks, in general, are supposed to be sparse. This is because, the sizes of the networks are huge, and the demographic variety exhibited in the network is unbounded. People from different geographical location, different ages, different socio-economic backgrounds, different interests and so on are all members of the OSNs. However, generally, people in OSNs connect to similar people, and thus, forms groups or clusters. Therefore, the number of connections over dissimilarities is very low. As a

result, the density in the small groups or clusters of familiar or similar people is high, but the inter-cluster density is very low. Also, the sizes of the clusters are considerably smaller than the size of the network. Therefore, in general, the density of an OSN is low.
The density for Facebook, YouTube, Email, Twitter, and Google+ are 0.0004, 0.0008, 0.0006, 0.00004, and 0.00007 respectively. The empirical study shows that the density of the OSNs is low, as predicted. However, the density of directed networks is in general lower than the density of the undirected network, by order of $10^{-}1$ or more. This is partially because the possible number in directed networks is twice that of undirected networks. Thus, even if the number of vertices and edges present in the directed network and undirected network is the same, the density of the directed network would be half that of the undirected network.

Global Clustering Coefficient

Global Clustering Coefficient of a network measures the presence of a particular microstructure, namely, a triangle, in the network. More specifically, GCC finds the ratio of triangles to triplets. A higher value of GCC indicates that most of the triplets are triangles. Triangle is the simplest representation of common neighbours, and thus, signify the formation of groups. Therefore, it can be anticipated that the value of GCC would be low for OSNs because OSNs have small groups and in general is sparse.
The global clustering coefficient for Facebook, YouTube, Email, Twitter, and Google+ are 0.1477, 0.0795, 0.0853, 0.0001, and 0.0008 respectively. It can be seen from the empirical study, that the value of GCC for the five OSNs considered for this research is low. However, the GCC of directed networks is lower than the GCC of undirected networks, by order of $10^{-}2$ or more. So, in directed networks, the presence of paths of length 2 is higher than the presence of cycles of length 3. Thus, the formation of groups is higher in the undirected network as compared to directed networks. This is partially because of the nature of the relationships and partially because of the purpose of the networks that have been considered.

Average Local Clustering Coefficient

The average local clustering coefficient of the network provides a general idea of the denseness of the network. ALCC measures how dense the local neighbourhood of a node is, for every node in the network. As the number of nodes in an OSN is high, and not all nodes are of the same importance, it can be predicted that the value of ALCC of a network should be low. For directed networks, since the neighbourhood of the nodes change with a change in the direction of the edge, the value of ALCC is different for in-degree distribution and out-degree distribution, which has been presented as $ALCC_{in}$ and $ALCC_{out}$ respectively.
The average local clustering coefficient for Facebook, YouTube, and Email are 0.2210, 0.1367, 0.49610 respectively. The average local clustering coefficient of Twitter and Google+ are $ALCC_{out} = 0.000002$, $ALCC_{in} = 0.0053$, and $ALCC_{out} = 0.000009$, $ALCC_{in} = 0.0549$ respectively. The observed values show that, as predicted, the value of ALCC is low for all the OSNs. However, the value of ALCC for directed networks is an order of $10^{-}1$ or more less than the value of ALCC for undirected networks. Thus, it can be

said that group formation is less in directed networks as compared to undirected networks as the neighbours of a node are not connected amongst themselves in directed networks.

Closeness Centrality

The closeness centrality value gives an idea of how closely connected a node is with the other nodes in the same connected component. So a higher closeness centrality value for a node means its average distance from all its connected nodes is low, that is, it is close to all its connected nodes. The values considered for the empirical study are the average closeness centrality for the OSN. The average closeness centrality of Facebook, YouTube, Email, Twitter, and Google+ as calculated for the empirical study are 0.2332, 0.2347, 0.2132, 0.0006, and 0.0006 respectively. It can be seen that undirected networks have a higher average closeness centrality compared to directed networks. This is indicative of closely knit groups in the undirected networks.

Eigenvector Centrality

Eigenvector centrality assigns importance to a node depending on the importance of nodes connected to it. Thus, if a node i is connected to important nodes, node i can also be considered as important. This measure can be significant in other applications such as information propagation. The eigenvector centralities observed for the five OSNs have been considered in the empirical study are the average eigenvector centralities. The average eigenvetor centrality of Facebook, YouTube, Email, Twitter, and Google+ as calculated for the empirical study are 0.0010, 0.0022, 0.0015, 0.0002, and 0.0033 respectively. It can be seen that the eigenvector centralities are similar for all the OSNs.

Betweenness Centrality

The betweenness centrality of a node measures what ratio of paths pass through that node, that is, the node falls between how many paths in the network. A higher betweenness centrality means a higher number of paths pass through that node. The observed values for betweenness centrality for the five OSNs considered in the empirical study are the average betweenness centrality for the network. The average betweenness centrality of Facebook, YouTube, Email, Twitter, and Google+ as calculated for the empirical study are 0.0001, 0.0002, 0.0001, 0.0000001, and 0.0000002 respectively. It can be seen that the average betweenness centrality for the directed network in order of 10^{-2} lower than that of undirected networks.

Homophily

Depending on the type of relationship of the network and the purpose of the network, the homophily value of an OSN will vary. For this research, the OSNs have been divided into clusters or groups using an algorithm proposed by Sen et. al.[45]. The clustering is done based on the local clustering coefficient values and the average local clustering coefficient value for the network. As the neighbourhood of a node changes with a change in the direction of the edge, the value for local clustering coefficient and average local clustering coefficient changes. As a result, the value of homophily also

changes. The average homophily value for in-degree distribution is presented as h_{in}, and that of the out-degree distribution is presented as h_{out}.
The average homophily of Facebook, YouTube, and Email are 0.5870, 0.6400, and 0.8237 respectively. The average homophily for Twitter and Google+ are $h_{out} = 0.0018$, $h_{in} = 0.0012$ and $h_{out} = 0.0027$, $h_{in} = 0.0021$ respectively. The observed values of average homophily inspire a few critical observations. Firstly, it can be observed that the homophily of directed networks in order of 10^-1 lower than the homophily of undirected networks. Thus, it can be said that groups with similar local clustering coefficients are not connected in directed networks. It can also be seen that there is a huge variation in the value of average homophily even for the undirected networks. This is partially because of the nature of the relationship and the purpose of the network.

Assortativity

The three possible levels of assortativity are - assortative, neutral, and disassortative. In assortative networks, that is, networks with assortativity value closer to 1, nodes with similar degrees are adjacent, and nodes with dissimilar degrees are not adjacent. In disassortative networks, that is, networks with assortativity value closer to -1, nodes with dissimilar degrees are adjacent, and nodes with similar degrees are not adjacent. In neutral networks, that is, networks with assortativity value close to 0, either both similar and dissimilar degree nodes are connected, or neither dissimilar nor similar nodes are connected. Therefore, depending on the nature of the relationship and the purpose of the network, the assortativity values of the OSN will differ. For directed networks, the assortativity of in-degree and out-degree networks have been represented as r_{in} and r_{out} respectively.
The assortativity value of Facebook, YouTube, and Email are 0.1770, -0.0752, and -0.1108 respectively. The average homophily for Twitter and Google+ are $r_{out} = -0.8812$, $r_{in} = -0.0534$ and $r_{out} = -0.3887$, $r_{in} = 0.0551$ respectively. The assortativity values observed for the five OSNs considered for this research lie within a small range of values close to 0. It can be observed that most of the networks behave close to neutral, with N_1 being slightly assortative and out-degree of N_4 being highly disassortative.

4.3.4 Analysis

It can be seen that observing the similarities and differences in the individual network structural properties is not enough. Although minor inferences about the structure of the networks can be inferred from these individual observations, it is not possible to analyse the overall structure of the network and the micro-stuctures of the network from these alone. For example, density, global clustering coefficient and average local clustering coefficient are all measures of network denseness. However, they can overlook important small structures present in the network. For example, in a network, there might be multiple small groups with high density or average local clustering coefficient. However, as the sizes of these small groups or clusters are many orders smaller than the size of the network, the presence of these structures cannot be inferred from the values of density of the network, global clustering coefficient or average local clustering coefficient alone.
Similarly, the variance in values of power law exponent, the values of d_{min} and γ all hint that the OSNs do not strictly follow power law. Also, the CDF plot of the degree distributions of the OSNs also hints at deviations from the power law. However, combining

the values of power law exponent(γ) and d_{min} along with the CDF plots of degree distributions of the OSN provides strong support that the OSNs do not strictly follow the power law.
Multiple inferences can be made about the network structure and micro-structures by collectively observing the values obtained for the multiple network structural properties. The value of Closeness Centrality, Density, GCC, ALCC, Homophily and Assortativity all show that undirected networks have more well-formed groups compared to directed networks. GCC shows that directed networks have more paths than cycles, disassortative/neutral nature of directed networks show that nodes with dissimilar degrees are connected. Therefore, it can be inferred that in directed networks, star-like structures are present with a high degree node at the centre. This inference is supported by the degree distribution of the directed networks as well. Similarly, closeness centrality, density, GCC and average local clustering coefficient of undirected networks are higher compared to directed networks. Assortativity and average homophily of undirected networks are also higher than directed networks. Thus, it can be inferred that undirected networks have small clique-like structures. This inference is supported by the degree distribution and local clustering coefficient distribution of the undirected networks. Also, considering the values of local clustering coefficient distribution, average local clustering coefficient, global clustering coefficient, density and homophily, it can be induced that the email network has a high number of well-connected cliques or pseudo-cliques.

4.4 Summary

This chapter provides an extensive empirical study of the network structural properties for the five OSNs that have been considered for this thesis. The results of the empirical study are recorded and analysed. The significance of the observed values have been explained in contrast to the expected values. The significance of the network structural properties in identifying the structure and microstructure of the network has also been highlighted. The results of the empirical study provides the basis for the categorisation of OSNs proposed in the next chapter.

Chapter 5

Proposed Categorisation of OSNs

The results of the empirical study shows that different OSNs vary with respect to their network structural properties. Based on the this, it can be inferred that the network structure and microstructures are different for different OSNs. From the empirical study it can also be seen that two network properties, homophily and assortativity, are sufficient for defining a categorisation of OSNs. This chapter proposes a categorisation of OSNs based on network structural properties. Fig. 5.1 represents the categorisation

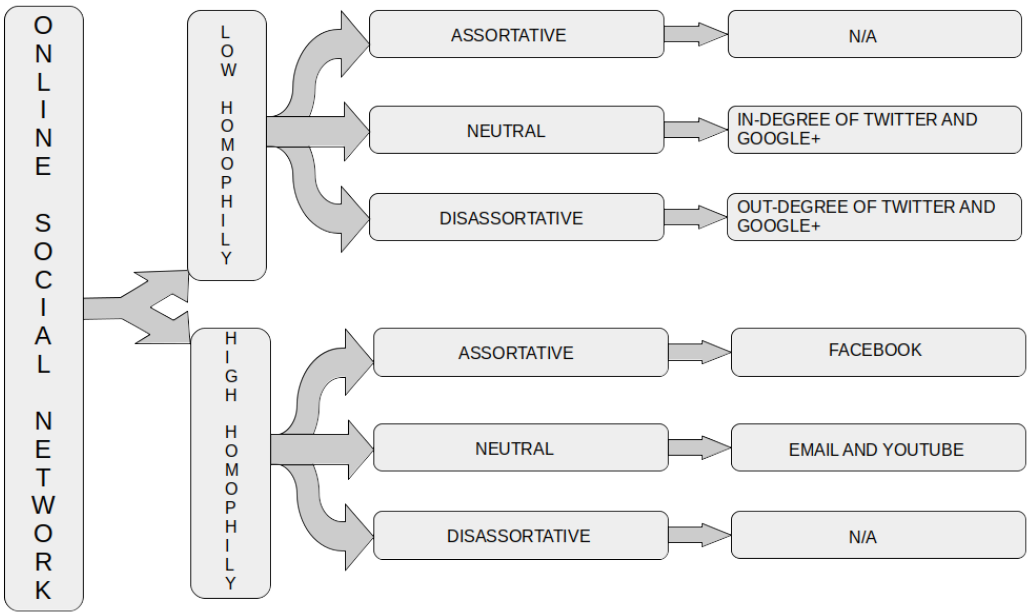


FIGURE 5.1: Categorisation of OSNs

obtained for the OSNs based on the observed values for different structural properties of the OSNs. As can be seen from Fig. 5.1, only two of the structural properties, namely, homophily and assortativity, were used for the categorisation. However, a similar categorisation can be obtained using a combination of other structural properties as well. Those categorisations have not been considered for this research as the categorisation represented in Fig. 5.1 is sufficient for the purpose of discussion.
The five OSNs that have been considered for this thesis have been categorised in four different categories. The categories are - Low Homophily-Neutral, Low Homophily-Disassortative, High homophily-Assortative and High homophily-Neutral. Detailed description and analyses of these networks have been provided below.

5.1 Low Homophily

5.1.1 Assortative

Low homophily signifies that the networks have less connection between nodes of similar clustering coefficients. Assortative networks are those where nodes with a similar degree are connected, and nodes with a dissimilar degree are disconnected. Thus, these two observations are contradictory and cannot co-occur. Therefore, no networks fall into this category.

5.1.2 Neutral

Neutral networks are those where either nodes with both similar and dissimilar degrees are connected, or neither nodes with similar nor dissimilar degrees are connected. Low homophily also indicates that the number of connection between nodes is low. Thus, it can be concluded that networks falling in this category have low density and low clustering coefficients as well. Examples of such networks are the in-degree and out-degree distribution of the directed networks Twitter and Google+. In these networks, as no particular shapes can be discovered.

5.1.3 Disassortative

Disassortative networks are those where nodes with a dissimilar degree are connected, and nodes with a similar degree are disconnected. Low homophily indicates that the nodes with similar clustering coefficients are not connected. Thus, it can be concluded that networks falling in this category will have very low density and clustering coefficients. Examples of such networks are the out-degree and out-degree distribution of the directed networks Twitter and Google+. In these networks, star-like structures are predominant as nodes with a dissimilar degree are connected, and the overall number of edges is meagre.

5.2 High Homophily

5.2.1 Assortative

Both high homophily and assortative signify that nodes with a similar degree and similar clustering coefficients are connected. Thus, it can be inferred that in networks of this type, the density and clustering coefficients will also be high. Examples of such networks would be Facebook. In this category of networks, the presence of pseudo-cliques is predominant. Because it is assortative, nodes with a similar degree are connected, and the high density and clustering coefficients indicate the number of edges in this category of networks is comparatively higher than the other categories, which suggests the presence of pseudo-cliques.

5.2.2 Neutral

High homophily indicates that nodes with similar clustering coefficients are grouped. Neutral networks, on the other hand, suggest that either nodes with both similar and dissimilar degrees are connected, or neither nodes with similar nor dissimilar degrees are connected. Thus, it can be said that networks falling in this category has average density and moderate clustering coefficients. Examples of such networks would be the Enron Email network and the YouTube network. Although these two networks have been grouped, they are mildly different concerning their clustering coefficients. The email network has higher average local clustering coefficient, higher global clustering coefficient and a higher number of nodes with local clustering coefficients in the range of 0.9 to 1. The YouTube network also has a lower number of connected components. However, the email network has a lower density than the YouTube network. Thus, it can be concluded that the Email network has a higher number of connected sub-graphs with a low number of nodes as compared to the YouTube network.

5.2.3 Disassortative

High homophily signifies that in the network, nodes with similar clustering coefficients are connected and grouped. Disassortative networks are those where nodes with a similar degree are disconnected, and nodes with a dissimilar degree are connected. Thus, these two observations are contradictory and cannot co-occur. Therefore, no networks fall into this category.

5.3 Summary

This chapter proposes a categorisation of OSNs based on the network structural properties. The five OSNs considered for categorisation in the previous chapter have been categorised into four categories. The expected network structure, microstructure and behaviour of the different categories of OSNs, as proposed in this chapter, has been explained. The next chapter concludes the thesis with remarks and future scope of research in related and relevant domains.

Chapter 6

Concluding Remarks and Future Direction

6.1 Conclusion

Online social networks have become a major part of the day-to-day life in modern society. User engagement is high on OSNs, and it results in the generation of much data. The user behaviour on OSNs is highly indicative of their behaviour in the real world. Just by studying the nature and structure of relationships on the OSNs, many inferences can be made about a user and the network as a whole. Thus, OSNs are no longer just a platform for connecting and communicating, but it is also a platform for marketing agencies and companies to showcase their products and services. OSNs have also gained popularity for participatory sensing, as a citizen reporting platform and for organising mass movements. Thus, OSNs have evolved into something much bigger than what they were initially envisioned as.
Under these circumstances, it is important to study the OSNs in details. There is a huge number of OSNs, many of which are very popular. They differ in purpose, nature and structure. Therefore, studying one OSN and generalising the results to all other OSNs is not feasible. Also, studying all OSNs individually is not feasible either.
Thus, it is important to categorise the OSNs and study the categories separately. If the categorisation is efficient, this provides a method of understanding the structure of a large number of OSNs without having to study each property in depth for each of the networks.
Based on the values of different structural properties, it is evident that the different OSNs are not similar with respect to network structure and microstructures. In-depth analysis of the similarities and differences in the structural properties of the OSNs yield the categorisation that has been represented in Section 5. The categorisation has been justified in the same section. It can, thus, be seen that it is a valid categorisation of OSNs based on structural properties.

6.2 Future Work

Proper categorisation of the OSNs opens up a wide array of research opportunities. This categorisation can be compared with other categorisations done previously, based on one or more of the structural properties, or based on the nature and purpose of the network. Similarities and differences in the categorisations can be analysed to get a better understanding of the networks.

The categorisation can be used to study different network phenomena for the different categories of the OSNs. Such phenomena may include the growth patterns for the different categories of the OSNs, the information diffusion and rumour blocking patterns for the different categories of OSNs and so on. Based on the results of these studies, new and improved category specific network models can also be designed.
The field of Online Social Network Analysis is still in its nascent stages, and a lot of open research problems are still present in this field of study. Proper categorisation of the OSNs can lead to significant progress in relevant research areas through a better understanding of the OSNs.

Bibliography

[1] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, vol. 13, no. 1, pp. 210-230, 2007.

[2] P. Erdos and A.Renyi, "On the Evaluation of Random Network", PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, pp. 17-61, 1960.

[3] Reka Albert and Albert-Laszlo Barabasi, "Topology of Evolving Networks: Local Events and Universality", PHYSICAL REVIEW LETTERS, vol. 85, no. 24, 2000.

[4] Duncan J. Watts and Steven H. Strogatz, "Collective dynamics of 'small-world' networks", Nature, pp. 393-440, 1998.

[5] Camelia Delcea and Ioana-Alexandra Bradea, "Grey clustering in online social networks", Vietnam Journal of Computer Science, vol. 4, no. 3, pp. 185–193, 2017.

[6] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan, "Clustering Social Networks", International Workshop on Algorithms and Models for the Web-Graph, WAW 2007: Algorithms and Models for the Web-Graph, pp. 56-67, 2007.

[7] Haris Memic. "Structural micro forces in online social networking websites: Impact on friendship structure", Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST 2009), IEEE Xplore, 2009.

[8] Albert-Laszlo Barabasi, Reka Albert, and Hawoong Jeong, "Statistical Mechanics and its Applications", Physica A, Elsevier, vol. 272, no. 1, pp. 173-187, 1999.

[9] Alan Mislove, Massimiliano Marcon, Krishna P. Gunmadi, Peter Druschel, and Bobby Bhattacharjee,"Measurement and Analysis of Online Social Networks", ACM Internet Measurement Conference, 2007.

[10] Edgar N. Gilbert, "Random Graphs", Annals of Mathematical Statistics 30: pp.1141–1144, 1959.

[11] Albert-Laszlo Barabasi and Reka Albert, "Emergence of Scaling in Random Networks", Science, vol. 286, no. 5439, pp. 509-512, 1999.

[12] Linton C. Freeman, "Centrality in Social Networks Conceptual Clarification", Social Networks, Elsevier, pp. 215-239, 1978.

[13] Linton C. Freeman, Douglas Roeder, and Robert R Mulholland, "Centrality in social networks: ii. experimental results", Social Networks, Elsevier, vol. 2. no. 1, pp. 119-141, 1979.

[14] Per Hage and Frank Harary, "Eccentricity and centrality in networks", Social Networks, Elsevier, vol. 17, no. 1, pp. 57-63, 1995.

[15] Tore Opsahl, Filip Agneessens, and John Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths", Social Networks, Elsevier, vol. 32, no. 3, pp. 245-251, 2010.

[16] Kazuya Okamoto, Wei Chen, and Xiang-Yang Li, "Ranking of Closeness Centrality for Large-Scale Social Networks", International Workshop on Frontiers in Algorithmics, FAW 2008: Frontiers in Algorithmics, Springer, pp. 186-195, 2008.

[17] Phillip Bonacich, "Some unique properties of eigenvector centrality", Social Networks, Elsevier, vol. 29, no. 4, pp. 555-564, 2007.

[18] K.-I. Goh, E. Oh, B. Kahng, and D. Kim, "K.-I. Goh, E. Oh, B. Kahng, and D. Kim", Physical Review E, 2003.

[19] Linton C. Freeman, Stephen P. Borgatti, and Douglas White, "Centrality in valued graphs: A measure of betweenness based on network flow", Social Networks, Elsevier, vol. 13, no. 2, pp. 141-154, 1991.

[20] M.E.J. Newman, "Mixing Patterns in Networks", Physical Review E, vol. 67, no. 2, 2003.

[21] M.E.J. Newman, "Assortative Mixing Patterns in Networks", Physical Review Letters, vol. 89, no. 20, 2002.

[22] Zhenkun Zhou, Ke Xu, and Jichang Zhao, "Homophily of music listening in online social networks of China", Social Networks, vol. 55, pp.160-169, 2018.

[23] Miller McPherson, Lynn Smith-Lovin, and James M Cook, "BIRDS OF A FEATHER: Homophily in Social Networks", Annual Review of Sociology, vol. 27, pp. 415-444, 2001.

[24] Kibae Kim and Jorn Altmann, "Effect of Homophily on Network Formation", Elsevier, Communications in Nonlinear Science and Numerical Simulation, vol. 44, pp. 482-494, 2017.

[25] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney, "Statistical properties of community structure in large social and information networks", Proceedings of WWW '08 Proceedings of the 17th international conference on World Wide Web, ACM, pp. 695-704, 2008.

[26] Ravi Kumar, Jasmine Novak, and Andrew Tomkins, "Structure and Evolution of Online Social Networks", Link Mining: Models, Algorithms, and Applications, Springer, pp. 337-357, 2010.

[27] J. Miller McPherson, Pamela A. Popielarz, and Sonja Drobnic, "Social Networks and Organizational Dynamics", American Sociological Review, vol. 57, no. 2, pp. 153-170, 1992.

[28] Gergely Palla, Péter Pollner, Albert-László Barabási, and Tamás Vicsek, "Social Group Dynamics in Networks", Gross T., Sayama H. (eds) Adaptive Networks. Understanding Complex Systems. Springer, pp. 11-38, 2009.

[29] Nicholas A. Christakis and James H. Fowler, "The Collective Dynamics of Smoking in a Large Social Network", The New England Journal of Medicine, 2008.

[30] Melanie E. Zaglia, "Brand communities embedded in social networks", Journal of Business Research, Elsevier, vol. 66., no. 2, pp. 216-223, 2013.

[31] Cynthia M. Webster and Pamela D. Morrison, "Network Analysis in Marketing", Australasian Marketing Journal, vol. 12, no. 2, pp 8-18.,2004.

[32] Robin Nunkoo, Dogan Gursoy, and Haywantee Ramkissoon, "Developments in Hospitality Marketing and Management: Social Network Analysis and Research Themes", Journal of Hospitality Marketing & Management, vol. 20, no. 3, pp. 269-288, 2013.

[33] Radhika Sharma, Vandana Ahuja, and Shirin Alavi, " The Future Scope of Netnography and Social Network Analysis in the Field of Marketing", Journal of Internet Commerce, vol. 17, no. 1, pp. 26-45, 2018

[34] Elder Semprebon, Danielle Mantovani, Rafael Demczuk, Cecilia Souto Maior, and Victoria Vilasanti, "Green consumption: a network analysis in marketing", Marketing Intelligence & Planning, vol. 37, no. 1, 2019.

[35] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua, "Attributed Social Network Embedding", IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 12, 2018.

[36] Ling Jian, Jundong Li, and Huan Liu, "Toward online node classification on streaming networks", Toward online node classification on streaming networks, Springer, vol. 32, no. 1, pp. 231-257, 2018.

[37] Shaojie Tang, Jing Yuan, Xufei Mao, Xiang-Yang Li, Wei Chen, and Guojun Dai, "Relationship classification in large scale online social networks and its impact on information propagation", 2011 Proceedings IEEE INFOCOM, 2011.

[38] Antonela Tommasel and Daniela Godoy, "A Social-aware online short-text feature selection technique for social media", Information Fusion, Elsevier, vol. 40, pp. 1-17, 2018.

[39] Sarbani Roy, Paramita Dey, and Debajyoti Kundu, "Social Network Analysis of Cricket Community Using a Composite Distributed Framework: From Implementation Viewpoint", IEEE Transactions on Computational Social Systems, vol.5, no. 1, pp. 64 – 81, 2018.

[40] Derek J. de Solla Price, "Network of Scientific Papers", Science, vol. 149, no. 3683, pp. 510-515, 1965.

[41] Albert-Laszlo Barabasi and Márton Posfai, "Network Science", Cambridge University Press, 2016.

[42] M.E.J. Newman, "The Structure and Function of Complex Networks", Society of Industrial and Applied Mathematics Review, vol. 45, no. 2, pp. 167-256, 2003.

[43] Stanley Milgram, "The Small World Problem", Psychology Today, vol. 1, no. 1, pp. 61-67, 1967.

[44] Agnieszka Rusinowska, Rudolf Berghammer, Harrie De Swart, and Michel Grabisch, "Social networks: Prestige, Centrality, and Influence", de Swart H. (eds) Relational and Algebraic Methods in Computer Science. RAMICS 2011. Lecture Notes in Computer Science, Springer, pp. 22-39, 2011.

[45] Faith Sen, Rolf Wigand, Nitin Agarwal, Serpil Tokdemir and Rafal Kasprzyk, "Focal structure analysis: identifying influential sets of individuals in a social network". Social Network Analysis and Mining, vol. 6, no. 17, 2016.